AI-Ready 材料数据、标准及基础设施

路勇超　汪洪　张澜庭　余宁

上海交通大学 材料基因组联合研究中心 和 材料科学与工程学院 上海 200240

摘 要 近年来，以“数据+人工智能（AI）”为特征的数据驱动新型研究范式得到了快速发展。围绕传统研究范式建立的数据产生、收集、存储、应用体系已无法满足新范式的要求，亟需建立以 AI 为导向的新型数据生态系统，以释放数据驱动的颠覆性优势。本文分析了 AI 方法的特点，针对性地提出 AI 语境下材料数据应遵循海量、全面、完整、均衡、可共享的原则。其中数据完整性和可共享性，是单条数据的特性，可以通过数据标准化得到保障。而数据是否满足海量、全面和均衡条件，更多地取决于数据生态的特点，需要全新的材料数据基础设施提供支撑。作为概念化的理想材料数据基础设施，“数据工厂”将颠覆现有的数据生产模式，带来材料数据数量和质量的全面提升，持续不断地提供 AI-ready 的数据。

关键词 材料基因工程，数据驱动，AI-ready 材料数据，数据标准化，数据基础设施

中图分类号 S220.4

AI-Ready material data, standards and infrastructure

LU Yongchao, WANG Hong, ZHANG Lanting, YU Ning

Materials Genome Initiative Center and School of Materials Science and Engineering,

Shanghai Jiao Tong University, Shanghai 200240, China

Correspondent: WANG Hong, professor, Tel: 17321042821, E-mail: hongwang2@sjtu.edu.cn
ZHANG Lanting, professor, Tel: (021) 54747471, E-mail: lantingzh@sjtu.edu.cn

Supported by the National Key Research and Development Program of China (No.2020YFB0704504), National Natural Science Foundation of China (No.52042301) and Shanghai ‘Science and Technology Innovation Action Plan’ Technical Standards Project (No.21DZ2206000)

Manuscript received 2022—**—**， in revised form 202*—**—**

ABSTRACT In recent years, the data-driven research paradigm featured by "data + artificial intelligence (AI)" has developed rapidly. The data generation, collection, storage and application system built around the current research mode can no longer meet the requirements of the new paradigm. Therefore, it is urgent to establish a new AI oriented data ecosystem to fully unleash the potential of data-driven paradigm. This paper puts forward that the principles which the material data for AI should follow, including massiveness, comprehensiveness, integrity, balance and shareability, based on the characteristics of AI methods. Among them, the data integrity and shareability, the characteristics of an individual piece of data, are ensured through data standardization. While the other requirements depend more on the data ecosystem, which requires a new material data infrastructure to support. As a conceptualized model of AI-ready material data infrastructure, Data Fab will revolutionize the

---

资助项目 国家重点研发计划项目 No.2020YFB0704504，国家自然科学基金项目 No.52042301，上海市科技创新行动计划”技术标准项目 No.21DZ2206000

收稿日期 2022-**-**，**定稿日期** 2022-**-**

作者简介 路勇超，男，1993 年生，博士生

通讯作者 汪洪，hongwang2@sjtu.edu.cn，主要从事材料基因工程理论、数据驱动的材料创新基础设施及数据标准体系研究
张澜庭，lantingzh@sjtu.edu.cn，主要从事材料基因组高通量实验技术及数据标准化研究

materials data production. It is conceivable that the Data Fab will serve as a reliable source for AI-ready material data and to bring about improvement in both data quantity and quality.

KEY WORDS material genome engineering, data-driven, AI-ready materials data, data standardization, data Infrastructure

2011 年，美国启动了材料基因组计划 (MGI)[1]，旨在利用定量数据和计算代码来发现和预测材料的行为，实现材料研发由试错法向预测型范式的转变，从而加快新材料的发现、设计、开发和部署，降低成本。其中心内容是发展先进的材料计算、实验和测试及数据信息学的工具，并将它们集成，构建新型的材料创新基础设施。欧盟、日本等发达国家也迅速启动了类似的政府主导的研究计划[2][3][4]。中国科学院和中国工程院自 MGI 发布的当年起，分别组织开展了广泛的咨询和调研。基于中国工程院关于中国版材料基因组计划咨询报告，汪洪等[5]对材料基因组的理念进行了归纳总结，并根据中国的实际需求特点与现有条件，对实施中国版材料基因组计划的发展战略、技术路线、政策措施等提出了建议。科技部于 2015 年启动了“材料基因工程关键技术与支撑平台”重点专项[6]。此后，汪洪等[7]进一步讨论了材料基因工程的三种具代表性的工作模式，阐明了材料基因工程方法与传统方法的根本不同点在于以数据为基础。明确提出数据驱动模式以“数据＋人工智能”为标志，围绕数据产生与数据处理展开，通过大量数据结合人工智能（Artificial Intelligence，AI）分析，揭示海量数据间的关联，挖掘潜藏的参量关系。数据驱动为材料研究开拓了新的视角，得益于 AI 的高效数据分析处理能力，数据驱动模式大幅度增加了研究问题的维度，提高了材料探索速度，从而有望带来颠覆性的效果。与之相比，实验驱动与计算驱动仍旧以传统的基于事实判断或理论推演的方式开展，并未改变既定的研究思维。因此，数据驱动代表了材料基因工程的核心理念与发展方向。近年来，以数据+AI 为基本方法的材料研究工作正呈快速上升趋势（如图 1），数据驱动的材料研究态势已初步形成。

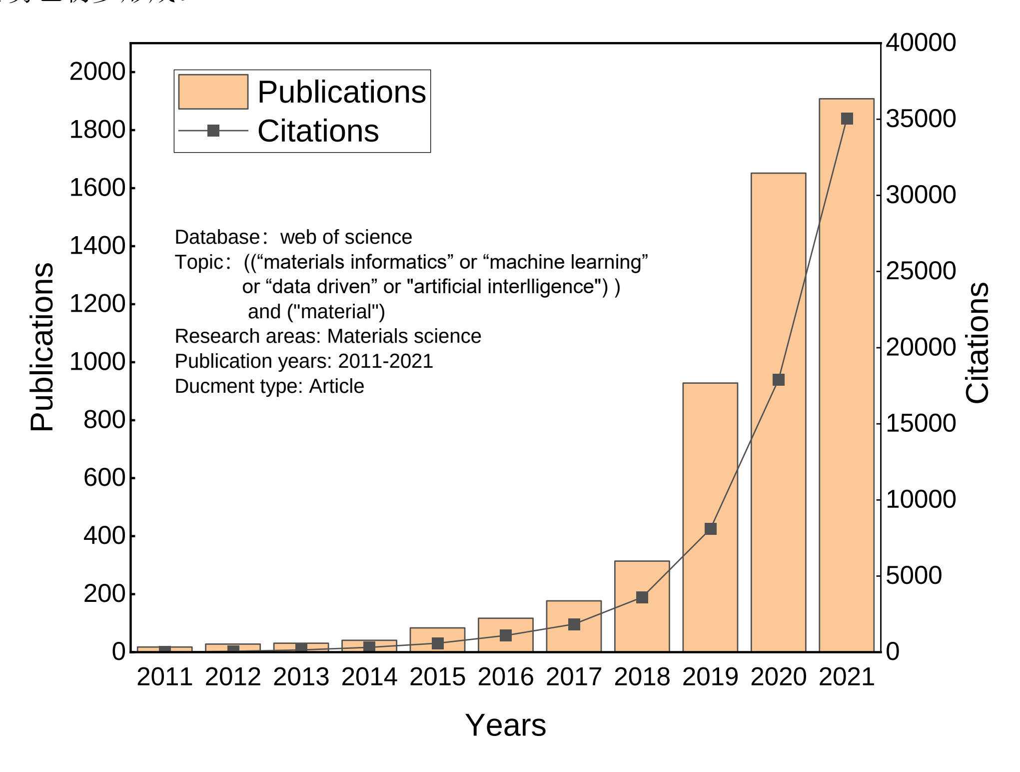


图 1 材料领域“数据+人工智能”研究的发表和引用趋势

Fig.1 Trends of publication and citation of "data + AI" research in the field of materials

关于人工智能的科学定义，可以从多个方面进行阐述[8][9]。从数据角度，人工智能（AI）被定义为“一个系统所具有的正确解读外部数据、从这些数据中学习、并通过灵活的适配使用这些习得知识来实现特定目标和任务的能力[10]”。这种能力为材料研究提供了一个通过数据间相关性来探索规律的方法。它有别于

传统物理模型所依赖的因果性，为缺乏基本物理模型条件下的科学规律研究提供了新的视角[11]。材料是由极大数量原子构成的复杂体系，材料性能经常是多个物理机制耦合的结果，很少只受单一因素影响，因此仅仅建立起性能与某一个参量相关的简单模型，很难描述清楚。从生活经验可知，通常人类大脑只能想象三维图像，同时处理超过三个变量以上的问题是具有很大挑战性的。利用人工智能方法可以轻松地同时研究成百上千个参量耦合的效果，这大大增加了理解问题的维度。因此，在解决这类问题时具有极大优势。与此同时，传统实验或计算研究所形成的先验知识在实际中常被用于为人工智能构建知识模型提供特征选择和模型优化、解释的基础参考[12][13][14]，因而，数据驱动并非是实验驱动与计算驱动模式的简单替代，而是在此基础上的补充和延伸。

人工智能基于数据而建立。数据的规模和质量与人工智能模型的可靠性呈正相关关系，因此，数据+人工智能共同构成了数据驱动范式的核心内容。简单来说，数据就是我们通过观察、实验或计算得出的结果[15]。在传统思维中，数据的主要作用是提供事实，作为科学研究、技术设计、查证、决策所依托的数值根据来使用，主要体现其表观价值。长期以来，材料科学数据生态是围绕计算、实验等传统研究范式而建立的。数据经常作为个体研究者在特定目标的研究中为获得特定信息所进行的实验或计算的结果而被产生并收集，因此整体呈现出多源异构、规模小、离散分布、无规范的特点。在人工智能背景下，数据是作为各种参数综合作用效果的承载体，为数据挖掘提供信息源。人工智能方法对大量数据进行处理与分析，通过建立数据间的关联，挖掘出背后构成这种关联的参数及相互关系，此时更多地是体现数据的内在价值。由于数据驱动模式在数据使用中表现出的新特点，对用于 AI 的数据在组织形式和内容上都提出了不同于往常的新要求。

目前已有的材料数据绝大部分都是面向传统的应用形式来收集、组织、存储和呈现的。事实上，这样的数据在基于 AI 的应用中，其查找、访问、准备、共享、重用和机器自动处理方面都遇到一定困难，这客观上阻碍、延缓了数据驱动模式在科学研究领域获得更快、更广泛地应用[16]。因此，在材料科学领域正阔步迈向数据驱动新未来的时刻，有必要对于在 AI 语境下材料数据应具有的特征、性质、特点取得深刻的理解与明确的认识，从而指导面向未来的材料数据的采集、组织、存储与使用，使之适合用于人工智能方法，助力其充分发挥出特殊的潜力。具有这样特点的数据在近期发布的新版美国材料基因组战略计划[17]中被恰当地称为 AI-ready（AI 就绪）。对 AI-ready 的含义做出清晰地解释将为构建面向未来的材料科学数据基础设施提出必要的基本遵循。这对于推动人工智能方法在材料科学领域中的应用，加速研究范式从试错法向数据驱动的预测型转变具有决定性意义。

本文从 AI 的自身特点出发，结合材料领域数据治理现状和最新趋势，对 AI-ready 的材料数据所需满足的特点、要求进行了全面分析，在此基础上进行总结，讨论了实现 AI-ready 的举措和领域内正在开展的相关工作。

1 AI-ready 对材料数据的要求

1.1 海量数据

人工智能本身融合了统计学的相关知识，需要有足够的样本量来表征所训练数据潜在规律的显著性[18]，再将其学习到的数据关联知识用于新样本的决策判断中。众多案例表明，随着模型训练集数据量的增加，模型愈加准确。例如 Schmidt 等[19]报告，钙钛矿化合物形成能的预测误差随着训练集数据量的增加呈幂指数单调下降，当训练集加倍大约可将误差降低 20%；Lee 等[20]研究了无机化合物带隙的机器学习模型，部分模型误差随数据量增加下降趋稳，而对于支持向量机模型，当达到该工作最大数据量时，误差仍呈明显下降趋势，说明数据量的增加将进一步促进模型的优化。因此，海量的数据是人工智能采用相关性策略探索的基本保障。

材料研究领域长期延续着课题小组的工作模式，研究社区主体以传统的低通量实验或计算方法来对材料的特性进行表征或模拟，再用产生的结果进行材料构效关系构建。工作模式的分散及表征、模拟方法的多样，造成材料数据来源众多，且研究社区内没有建立明确而统一的数据管理规范，导致了各个研究团队采集数据的种类和格式互不相同，数据呈现出多源异构的特点[21]，即便以某个具体材料类型为主题来汇集

研究数据，比如镍基高温合金材料的制备、表征数据，因为不同团队数据模板格式的差异性，总的可用数据量仍旧不会太大[22]。可用数据匮乏问题在机器学习相关的研究工作和评述文章里经常被提及[23][24][25][26]。

目前材料领域对海量数据获取途径可大致分为两种：

（1）高通量实验与计算技术，是高效产生大量材料数据的直接手段[27][28][29]。例如，以组合芯片为代表的高通量制备技术[30]，可在一块1英寸见方的基板上快速制备包含覆盖完整三元系成分含量的薄膜样品。采用同步辐射微束X光面探衍射技术对其进行表征，单点衍射表征时间可缩短到1-2秒，在一块组合材料芯片样品上获取5000点以上的衍射谱图，总耗时在7小时以内，单日可完成3块组合材料芯片的逐点结构表征工作。以第一性原理计算为代表的高通量计算依托先进超级计算机的超强算力、智能纠错的自动化计算流程、规范化的计算参数设定，可高速批量化地产出大量服务于材料设计的计算模型数据[31]。高通量实验与计算技术是从根源上加快产生材料数据量的有效方式，我国在十三五期间通过材料基因工程重点研发计划专项对高通量实验与高通量计算技术进行了系统布局，并取得了诸多进展[6]。各个细分材料领域正在持续推进该工作的开展[27][32][33]。

（2）从海量文献提取数据[34]。迄今为止，各种公开发表的科学文献是大量重要的科研数据的最主要出口与聚集地，将它们收集起来具有重要意义[34]。目前研究成果的呈现并无标准形式，大部分均以非结构化的异构形式公开。Pauling File 项目[35]是最大的人工收集无机晶体材料数据的项目之一，收集了从1891年至今材料科学、工程、物理和无机化学的科学文献中提取的晶体结构、物理性质和相图数据，迄今总共包含了超过350000个晶体结构、150000个物理性质和50000个相图，并于2016年推出了在线版本MPDS(Materials Platform for Data Science，https://mpds.io)。同时，借鉴生物医药信息学领域的经验，研究者们开始尝试采用自然语言处理、文本挖掘方法等计算机技术来自动化地从文献中提取数据。英国剑桥大学J. Cole开发了一个用于化学文本的自然语言处理工具包ChemDataExtractor[36]，并使用它构建了磁性材料相变温度的大型数据集[37]，以及电池材料电化学性质的数据集[38]。从文献中提取数据是对当前非结构化数据发表生态的一种弥补性方案，其中手动提取模式需要专家知识来进行标注，数据精度较高，但耗费大量人工，效率较低；而采用自然语言处理和文本挖掘算法来自动提取文献数据效率比较高，但是精度比较低。从文献中提取数据是一种间接的数据收集方式，数据在非结构化发表和再次抽取的过程中，会导致大量有效信息损失。因此有必要改革知识确权方式与共享机制，将有价值研究数据直接发表。

1.2 综合全面的特征量

材料数据中所包含的特征量决定了AI描述现象的可能视角。如果数据中仅包含单一特征量，由此产生的认识必将局限于研究变量与此特征量的相互关系，而无法延伸至除此特征量之外的特征。一套能完整反映材料研究过程的特征集将有助于AI对数据间关联产生更精准的认识。在传统研究模式中，由于人类生活在三维空间中，人脑仅可直接处理较低维度的研究问题，在科学推理时，经常采取理想化的形式来对自然现象进行简化，比如经典物理学中经常用“足够光滑平面”、“忽略空气阻力”、“理想气体”等理想化假设，去掉一些复杂的干扰因素，只保留关键因素进行研究分析。在现代材料科学所采用的探究工具中，受限于技术条件，也经常采用一些类似手段来保证科学探究的可开展，比如“真空条件”、“模拟海水腐蚀”等。这些简化反映在数据上，就是对复杂高维的现象经过降维进行低维描述，以方便人类对其进行处理。当然，也不可避免地引起导致了真实世界与认识的一定偏差。

人工智能方法的特点之一便是有能力处理高维度的数据，这为探究认识更真实的自然世界提供了新途径。与此相适应，AI-ready的数据集应包括尽可能综合全面的特征参量，以充分发挥人工智能的潜力。从工作流程上看，实验驱动和计算驱动均是先提出可能的理论，再搜集数据，并通过表征或仿真方法进行验证。这种依赖先验知识的做法有利于聚焦已知特征参数进行高效优化，但受限于当时的认识水平，有可能在无意中排除掉许多可能在实际问题中同样有意义的参数。从而在实际工作中限制了我们的想象力[39]，导致一些未知的关键因素擦肩而过。而数据驱动范式从理论上说，并不预设哪些参数是重要或不重要的，也就避免了对参数选用的习惯与偏向。例如Ward等[40]围绕化学元素的计量属性、统计属性、电子结构属性、离子化合物属性等四个方面，创建了一组包含了145个材料参数的通用性特征空间，可对任意化学元素组成的无机材料进行特征表示，结合各种机器学习模型和训练数据，能够对材料的物理、化学性能进行预测，并在晶体的带隙能量、比体积、形成能预测和新型非晶体的发现两个不同方面上验证了其通用性和有效性；

同时，为了量化每个特性对目标性能的预测能力，依次采用二次多项式拟合方式来测量模型的均方根误差，发现对于不同的材料和性能，影响其最佳建模的特性参量可能会发生显著变化。比如金属间化合物的形成能与熔化温度的变化和组成元素之间的 d 层电子数最相关，而含有至少一种非金属的化合物与平均离子特征（基于组成元素之间电负性差异的量）关系最密切，这些示例中最相关特性的变化进一步支持了构建机器学习特性集中有大量可用特性的必要性。该工作中所涵盖的 145 个特性虽然无法完全涵盖无机材料的所有特征，但朝着创建丰富的材料特性空间迈出了一大步，体现出全面的特性空间对于人工智能自动分析探索，获取未知规律的重要价值。

1.3 数据记录的完整性

从材料研究数据的产出过程来看，这些数据中不仅仅揭示了材料样品自身的内在特性，也蕴含了材料的制备、表征、计算设施及处理流程等相关因素的影响[41]，利用 AI 对研究数据分析处理时，这些因素均将在数据所反映的内在关联关系上有所体现。在以工艺优化、性能改进等为目的的研究中，研究者能够有效的获取、利用这些隐藏关系的前提是数据集中包含可反映制备、表征、计算等研究过程的完整特征维度，才可在相应特征参数上才能得到精细化、定量化的参考指导，并在计算模拟和实验中快速实现。

同时，任何制备、表征、计算过程都包含了大量细节参量，AI-ready 的数据必须对这些参量有足够完整的收纳，使数据使用者对数据产生的条件、环境、过程充分理解，如同他们自身经历过一样，才能真正确保对这些数据的正确、合理使用。从当前的数据采集方式看，数据产生者主要是基于自身的研究目的来进行材料的实验制备、表征或计算模拟研究，记录每一条数据时，往往仅选用一部分符合自身研究需求的"关键参数"，而将研究过程中产生的其他参数直接忽略或舍弃。这些"不完整"的数据记录经过发表被再次使用时，常常由于细节的缺失而导致研究结果不可重复，这在各个科学领域都经常出现[42]。为了保证科学工作的可靠性和科学数据的可重用性，目前一些期刊开始要求用户在提交预发表文献的同时，需要同时提交该成果中的所有源数据，如 Nature 出版社旗下的所有期刊都有该要求[43]，并鼓励作者将所有必要数据存储在公共存储库中公开，并描述数据获取的完整途径，一些推荐的公共存储库包括 Figshare[44]、Zenodo[45]和 Dryad[46]等。考虑到数据驱动范式下数据使用时空范围在不断扩大，不同使用者对数据的利用视角也愈发广阔，在进行原始数据采集时，需充分考虑对数据产生动作相关的参数做到"应收尽收"，留下完整的数据参量记录，为数据的再利用提供尽可能详尽的信息，并为 AI 高效指导材料的优化设计提供详尽、全面的特征空间。

1.4 数据分布的均衡性

如前所述，人工智能通过找出多个参量间相关性来揭示数据内在规律。然而若用于训练的数据集在参数空间分布不均衡，将导致标准模型的判断结果发生偏差[47]，这在 AI 应用较为成熟的商业系统应用中较为常见，比如亚马逊曾放弃了一个通过 AI 来对求职者简历进行评分的智能招聘系统，因为该系统对女性应聘者产生了不公正的判断结果，出现这种偏见的原因是用于开发算法的训练数据集是基于与以前的申请人（主要是男性）相关的数据[48][49]。

类似地，当 AI 用于材料科学探究时，若材料特征数据集中带有人为的偏向，将导致模型的判断结果也出现相应的偏向，如在传统方式的材料科学研究中，研究者往往只注意记录与研究目标相符的所谓"积极数据"，而将与目标不符的"消极数据"直接忽视或舍弃，这样收集到的数据用于人工智能模型训练时，会导致模型在统计意义上丢失部分客观性，并会损失一些潜在材料规律的挖掘机会。本质上，科学数据的所谓"好坏"是研究者从狭义角度进行的人为定性，数据本身是无优劣之分的，从统计学角度看，在严谨的科学条件设计下，每次材料实验产生的数据都是对材料客观规律的一次反映，均应该进行记录保存，在后续利用 AI 进行数据分析时，模型才能够全面客观的反映材料规律，具备较强的鲁棒性和可扩展性，充分体现出每条数据的潜在价值。例如在 2016 年发表的著名案例中，Raccuglia 等[50]在使用决策树方法预测新的金属有机氧化物材料时，在训练集中同时包括了之前的"成功"与"失败"的实验数据。

1.5 数据的可共享性

科学数据的重复利用，是关系到科研文化由单打独斗向共享合作的大科学模式改变的根本性要求，也是大数据时代数据驱动模式的现实需求。为构成所需的海量、多参量、均衡分布的数据集，单一来源数据

往往很难满足，需要将多个来源的离散数据整合。这就要求离散的单条材料数据在形式表达上具备参与到大数据集的条件，满足使用者对数据便捷访问、使用的可共享需求。

近几年，数据共享受到了广泛的关注，我国十三五材料基因工程重点专项也对数据汇交做出专门规定，提出硬性要求。美国 2019 年发布的《联邦数据战略和 2020 年行动计划》[51]、欧洲 2020 年发布的《欧洲数据战略》[52]、中国 2018 年发布的《科学数据管理办法》[53]等，均从国家战略层面制定了促进科学数据共享的配套政策和实施方案支持。在一些具体科学领域，也部署了促进数据共享的强制性措施，比如美国国立卫生研究院 (NIH)规定[54]，自 2023 年 1 月起，将要求其每年资助的 30 万名研究人员和 2500 家机构中的大多数在其拨款申请中纳入数据管理计划，并最终公开其研究数据。数据共享已在科学界形成共识。

然而，一直以来，科学数据大多存储在本地服务器上，且缺乏明确一致的管理规范，不同来源的数据在表达格式、表述完整性上参差不齐，使得数据既不容易访问，也不容易集成利用，共享效益较低。围绕科学数据如何能被更广范和更充分的利用这一问题，国际科学界已经探讨多年[55][56]，2016 年，荷兰莱顿大学的 Barend Mons 教授联合学术界、产业界、资助机构和学术出版商等一系列数据利益相关行业的代表，共同设计认可了一套简明且可衡量的数据管理原则——FAIR（Findable（可发现）, Accessible（可获取）, Interoperable（可互操作）, Reusable（可再利用））原则[57]，用于在更广范围提升数据的可共享性和可重用性。FAIR 原则得到了科学界的广泛认可，一些新型数据共享基础设施建设正在基于 FAIR 原则进行建设[58][59]。FAIR 原则的基本要求可以总结如下：

可发现（Findable）原则针对 AI 研究所需的目标特征研究数据从何处查询、用什么来查询的问题提出，要求数据被唯一且持久的标识符进行标识，其典型代表为 DOI 系统（Digital object identifier system）[60]，能够在公共互联网空间通过系统分配的唯一数据标识，为数据对象提供一种解析访问方法，找到目标数据的所在存储位置；同时要求用丰富的元数据来描述数据，并将其在可检索的数据资源平台中注册或设置索引，使查询者能够通过数据的特征属性来对目标数据进行精确检索，满足了目标数据能够被使用者查询到的基本要求。

可获取（Accessible）原则对于所查询到的目标数据如何获取的问题，对数据的获取方式进行了最低实现程度的规定，要求在开放、免费、可普遍实施的标准化通信协议下检索数据及其元数据，使得数据能够通过网络基础设施免费、简捷的进行传递；涉及到数据获取过程中的知识权益问题，允许数据所有者设置数据获取权限，在必要时对数据使用者进行身份验证和授权流程，在尊重数据所有权归属的基础上鼓励数据的开放；同时要求描述数据基本信息的元数据应能够持久访问，即便数据对象因为各种原因变更而不可访问，保证数据所携带的信息能在最低限度上被稳定的获取到；此外，可再利用（Reusable）原则要求在数据发布时应包含清晰且可访问的数据使用许可要求，为数据能够被正确的访问获取提供明确的注解提示。

可互操作（Interoperable）原则要求数据使用正式、可访问、可共享和广范适用的语言来描述数据，其中所涉及到的词汇应从符合 FAIR 原则的词汇表中或已有的权威术语中来选择，从而使其表达形式在领域内具有通用性，避免不同来源数据集成时在数据语义、格式上的不兼容，使得不论是人类或机器均可方便的对数据进行处理，为 AI 应用建立一套领域共识的可理解语言机制。

对于非自身产生的数据，在 AI 模型构建时如何完整理解和正确使用这些数据的问题，可再利用（Reusable）原则要求用多个准确且相关的元数据来描述数据，这些元数据应与数据的详细出处相关，且在表达组织上符合本领域相关的标准，使得使用者能够尽可能详细的了解到数据的背景和内容组成，促使其能被合理利用。可再利用原则充分考虑了非数据产生者在完整理解数据时所应具备的内容要求，为 AI 模型所需多来源数据的正确理解、使用和模型解释、应用提供可靠性保障。

2 AI-Ready 材料数据的实现

2.1 材料数据的标准化治理

AI-ready 对材料数据的海量、全面、完整、均衡、可共享需求，反映了数据驱动研究范式下的新型数据生态特点。其中数据完整性和可共享性，是单条数据的特性，可以通过标准化方式得到保障。标准化是为在既定范围内获得最佳秩序,促进共同效益,对现实问题或潜在问题确立共同使用和重复使用的条款以及编制、发布和应用文件的活动[61]。现实中的材料数据覆盖材料研发的全链条，从电子、原子、分子现象，

多尺度下工艺条件对材料性能与服役表现的影响，直至应用设计与制造技术的细节。管理这样海量且多元的数据需要全领域的协调建立共同的规则，从而无缝地实现数据的交换与共享，实现 AI-ready 的目标。传统材料数据库一般收集由原始数据处理而得到的分析结果（如各种材料性能参数等），而原始数据通常分散在实验者手中，不被收录，且数据格式五花八门，不便为其他人再次利用。再有，这些数据产生时往往以特定应用为目标，包含的材料属性相对有限，缺乏综合性。这样，数据可关联的参数就比较有限。这与传统材料研究方式与数据产生方式有着极大关系。因此现有的材料数据库大多不能满足材料基因工程的需要。在数据驱动前提下，有必要通过顶层设计，提出建立符合 AI-ready 要求的材料数据结构的通用规则，用于规范 AI-ready 数据的内容组成。

数据标准为 AI-ready 数据库（集）的构建提供了重要的保障措施。材料数据具有数量大、种类多、形式多样、产出单位各异、知识产权归属复杂等特点，如果没有统一的标准可以遵循，不仅收录存储更加复杂，也不便使用。在当今多种数据基础设施共存的条件下，某种形式的标准化是实践数据驱动范式所必不可少的[62]。因此建立统一的数据标准是围绕数据的规范化治理所开展的关键措施，为材料领域大规模采用人工智能方法奠定重要基础。

2.1.1 AI-Ready 数据标准化的内容

元数据是一种较为直观的数据组织管理方式。元数据通常被定义为关于数据的数据，本质上是从某个角度对数据对象进行结构化描述的一种形式。例如对某个人进行描述可以通过姓名、性别、身高、年龄、性格等众多元素进行描述。从特定角度反映数据对象所具有的特征，需要选择相关元素组合形成特定的元数据模式。标准化就是以在一定社会范围内取得共识的方式来规范元数据模式中所涵盖的内容。在数据驱动模式下，元数据是数据检索和人工智能分析的实际载体。数据的完整性和可共享性可以通过在标准元数据模式中包含相应的元素来得到保证。由于材料体系复杂，种类众多，为材料科学开发信息丰富、详尽且适应性强的标准化元数据是一个突出的挑战[62]。目前在材料元数据标准建设方面，国际上尚处于起步阶段，现有的元数据标准不是完全缺失就是不完整，标准组织(如国际标准组织（ISO）)为提供受控词汇表、数据格式和数据处理等元数据规范化相关的标准进行了许多尝试，但到目前为止还没有在领域范围内得到采用[63]。

本体（Ontology）是对“共享概念模型明确的、形式化、规范化说明”[64]（An ontology is a formal, explicit specification of a shared conceptualisation）。本体能够描述某个领域内的特定概念体系及其中各元素之间的确定关系。在实际构建层面，本体本身并没有定义其表现形式，可通过 OWL、DAML、RDFS、IDEF5 等多种语言表示[65]，将本体设计转化为计算机可处理的模式，目前较为常用的本体语言是 OWL 语言（Web Ontology Language）。各类本体在表达结构上具有相似性，均采用概念（也称为类）、实例、属性、关系、约束等基本构造元素来进行更具体的描述[65]。举例来说，我们在描述 45#钢材料“抗拉强度”和“延伸率”等数据“属性”时，这个概念体系包括：“材料”是一个“类”，代表所有类型的材料；在“材料”中还可分“金属材料”、“无机材料”、“高分子材料”等子类（“子类”代表了它们之间的“关系”），“钢铁材料”又是“金属材料”的子类；“45#钢”是“钢铁材料”中的一个具体实例；这个实例具有“抗拉强度”和“延伸率”等多种“属性”，而抗拉强度 460MPa，延伸率 17%”定义了“45#钢”这个实例的两个属性值。再有，“铁素体钢”可以定义为一种包含至少一种铁素体组织的钢，我们可以用材料本体中的“钢”、“铁素体”和“基本组织”之间的关系来约束定义“铁素体钢”这一概念[66]，这种约束令数据对人类和计算机均有意义，是建立计算机对概念体系进行自动推理的基础。本体与元数据均是描述数据资源的工具，二者均通过概念，或者说术语来对对象所包含的特征进行表示，区别在于元数据通过树状形式来组织这些术语，在表达上更加模块化和直观简洁；而本体通过网状形式来表示，更加凸显这些术语的相互联系，为数据资源的理解和利用提供语义背景。它们之间可通过其所包含的术语及其关系进行相互转化。标准元数据模式可以被表达为在一个学科范围内定义的一套表述规范、相互关联的一个通用概念体系，其中的元数据元素根据其在概念体系中的逻辑关系，可以看成是体系中不同概念的构造元素，因此本体可以用于描述、反映元数据元素间的关系。

在材料科学中，现有材料本体是对材料、材料性质、单位和约束条件及其相互关系的一种分类方案[62]。标准化材料本体的建立将为材料领域内研究者提供一个共享的标准概念体系，促进领域内不同研究者对同

类数据描述、管理的规范化协同，提升数据间的互操作性。多个异构数据库之间的数据交换可以方便地通过基于材料本体的中间数据表示来实现。随着本体的采用范围扩大，还将释放机器自动推理、挖掘海量材料数据间所隐含知识关联的潜力。目前，有关材料科学的本体建设刚刚开始，距离覆盖完整的知识体系还有很大差距。同时，各种各样的本体和不太正式的标准相互竞争[62]。如 NOMAD Meta-info[67]、ESCDF[67]和 OpenKIM[68]是原子材料科学中对计算结果进行分类的初期尝试，PLINIUS[69]用于陶瓷领域，ONTORULE[70]用于钢铁行业，SLACKS[71]用于层压复合材料，PIF[72]、Ashino[73]、EMMO[74]、MatOnto[75]、Premap[76]和 MatOWL[77]代表一般材料科学数据，等等。还没有出现确保材料完整表示的标准化本体。虽然材料本体的发展过程已经加快，但它们还没有像其他领域（如生物科学）那样成熟[66]，在工业应用中，这些公开可用的本体通常是不够用的，这迫使商业公司创建自己内部的、特定使用范围的本体[74]。

2.1.2 材料数据标准的国内外现状

近几年，材料信息学领域已经开始广泛认识并重视数据标准化的重要性[6][41][62][63][67][78]。但在实际操作中，建立并推广标准是一件耗时费力的工作。尤其是在数据库基础较好的国家，形成各家共识本身就似乎是一件不可完成的使命。为了应对快速积累大量数据的需求，以美国国家标准与技术研究院（NIST）为代表的机构采用了数据仓库的做法，即不限制材料数据的格式，将数据尽量多地收纳存储起来，以待今后进一步开发工具进行整理、分析和挖掘。数据仓库的形式对于解决数据量瓶颈问题是个短平快的方案，同时也是对缺乏数据标准现状的一种妥协。随机技术和标准的进步，后期固然可以对数据做一些标准化的规整，但原始数据中本身缺失的信息是无法通过事后弥补的。因此需要尽量一开始就标准化。

欧美国家注重对既有数据与数据系统的利用，尽力通过建立整套材料科学本体，改善多源异构数据的可互操作性，但这种元数据协调方式仍需开发数据转换器和共享数据模式。欧洲的新材料发现（NOMAD）实验室专注于收集、存储、清理计算材料学的数据，例如他们可以直接存储世界上主流的 10 多种从头计算代码产生的原始数据，然后通过开发翻译器的方法将原始数据规整为符合一定标准的格式[67]。FAIRmat[63]是德国国家研究数据基础设施（NFDI，https://nfdi.de)支持建设的数据联盟组织，将为材料领域许多特定的数据存储库构建一个联合基础设施，所有参与的团体或机构将使用统一的框架管理其数据，即在计算、管理和存储中共用一个中央元数据存储库。由于不同子领域不同主题的元数据存在差异，在管理时采用自下而上的分层方式，提取其共性元数据元素到上层中作为公共属性，比如材料的成分及研究方法，由此形成一个类似购物网站似的层层递进的数据组织和查阅模式，基于这些元数据形成一个材料数据描述的百科全书，可同时支持非专家用户的普通查询和专家用户的特定需求查询。FAIRmat 已经开始为不同领域中使用的词汇的数字翻译建立元数据和词典，下一步是开发本体，建立元数据之间的上述层级及其他关系描述，之后将标准元数据和本体部署到电子实验室记录本（ELN）和实验室信息管理系统（LIMS）中，实现不同团体所采集和存储数据的可互操作性。这种自下而上的元数据规范化工作模式使得 FAIRmat 在连接新的子域时具有较高的灵活性，但这种元数据协调方式需要开发数据转换器和共享数据模式。这种元数据协调方案的一个具体例子是 Open Databases Integration for Materials Design （OPTIMADE）[79]联盟最近发布的首个版本 API，通过该 API 允许用户访问参与该联盟的各数据库元数据模式项的公共子集，实现对分布式数据库的统一访问。

我国在开展材料基因工程方法探索与研究的早期，便认识到标准的重要性。2017 年在中国材料试验标准委员会（CSTM）成立之时，我国科学家前瞻性地便提出成立 CSTM 材料基因领域标准化委员会，这是国际上第一个材料基因工程领域的标准委员会，率先开展材料基因工程领域标准与标准化的重要探索与示范。2017 年 11 月 22 日，在第一届材料基因工程高层论坛期间，CSTM 材料基因领域标准化委员会（CSTM FC-97）正式成立，下设通则、计算、制备、表征、数据、应用 6 个技术委员会，分别负责对材料基因组的研究、开发、应用等各领域的材料产品、材料工艺方法、材料试验方法、材料试验技术评价方法、材料评价方法、材料模型和软件、材料计算、材料数据规范、材料领域管理和工作标准的团体标准体系建设工作。

考虑到材料基因工程以数据为核心的特点，FC-97 委员会确定将材料相关标准制定围绕数据展开。目前国际上尚无现成的材料基因工程数据标准可以借鉴。参考国际上材料数据标准建设中的实际情况，并结合中国材料研发领域特点与制度优势，FC-97 提出的标准化总体建设方针是：通过顶层设计，建立一个面

向未来的、适合材料基因工程数据系统的数据标准体系。数据标准体系中将包含一系列标准与规则，覆盖材料数据的全生命周期各个环节所涉及的技术、流程与功能，规范数据条目必须收集的内容与遵循的格式、协议、规定，使获得、存储与使用的材料数据都满足 AI-ready，符合数据驱动模式的要求。

首先，FC-97 选择在 CSTM 平台上从建立数据通用规则入手，基于最大化满足数据的 FAIR 原则这个基本出发点，确立数据条目中所包含内容的原则。2019 年 8 月，CSTM 发布了由国内 30 余家材料研究主体单位共同制定的世界范围内首个关于材料基因工程数据的团体标准—T/CSTM 00120《材料基因工程数据通则》（简称《通则》）[6]，《通则》跳出了材料及分工多样性对标准工作开展的限制，从数据层面切入，提出一套兼容性极强的材料数据分类框架。如图 2 所示，《通则》针对材料科学在数据驱动模式下对数据的需求，将数据分为样品信息、原始数据（未经处理的表征数据）与衍生数据（经分析处理得到的数据）三类，这里，样品可以是实验产生的实物，也可以是经计算产生的虚拟物。同理，原始数据可以来自于表征或是直接的测量，也可以通过模拟计算产生。注意，这里每条数据以单个操作（样品制备/表征/计算/数据处理）为单位，仅收集与该操作相关的内容。例如，关于样品信息的一条数据中只包含关于该样品制备的信息，而不包含对该样品进行表征的内容。对每条数据分别赋予独立且永久资源标识（例如依据国标 GB/T 32843 等规则，也可依据任何独立赋予的唯一且永久的标识体系）。

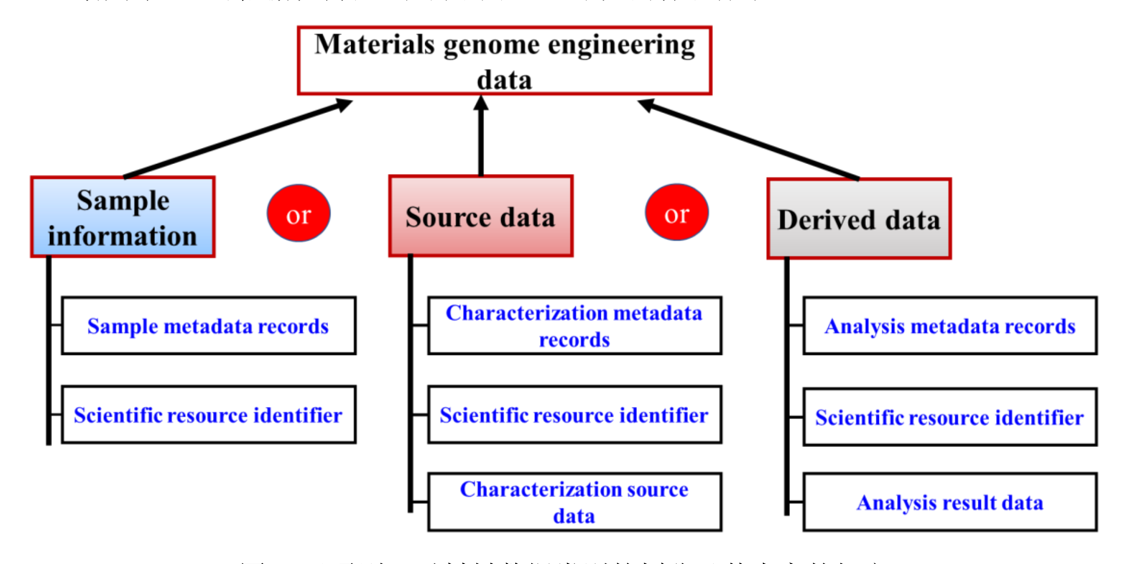


图 2 《通则》对材料数据类别的划分及其内容的规定

Fig.2 The classification of material data categories and their contents in the *General rule for materials genome engineering data*

《通则》的设计重点解决三个问题：其一，原始数据（未经分析处理的数据）中包含大量的信息，它的多次利用，特别为不同的目的多次利用是数据可再利用性的重要保障。目前原始数据大多分散在产生者手中，不被收录，极大地限制了数据的再利用。这样的分类从制度上确保原始数据被记录下来，从而保证了被再次利用的可能。其二，传统数据目前以数据产生者视角将成分-结构-工艺-性能间关系一体式组织呈现，从形式上就限制了数据应用的领域范围，不利于应用面开拓。《通则》将数据条目内容单元定为单个动作（制备/表征/处理），在保障丰富的元数据前提下，单条数据可依据自身信息独立的流通使用，方便地参与到使用者多元视角的材料探究中，在不同研究目的、情境下灵活自由的组合、重复使用。

其三，将样品单独列为一类数据是之前任何其它数据中都没有的做法。这样做的最大优点是使样品本身成为符合 FAIR 原则的公共社会资源，便于样品以数字代理形式共享、多用和重复使用。除此之外，还有以下几点重要考量：1）避免在表征元数据和衍生数据中包含过大且重复的样品信息所导致的数据处理负担，特别是衍生数据中可能大到不可接受；2）样品单独立项的前提假设是每个样品都是与众不同的个体，即便是两个表观参数完全相同的样品，其反映的重复性在材料数据科学中是具有统计意义的。传统数据库以一个样品作为同名样品的代表，实际上假设了所列参数是给定材料的特征值，客观上抹杀了由细节因素带来的差别。

目前，基于《通则》原则的材料基因工程术语标准、数据标识标准、数据通用规范等一系列规则性通

用标准正在建设过程中[80][81][82]，分别为各类研究方法的数据标准制定提供权威术语、标识方法、标准化流程与方法参考，以更具体的服务数据标准化工作建设。例如，充分的元数据是数据再利用的基础条件，是 AI-ready 要求中的重要组成部分，目前材料数据收录的元数据通常很不完整，达不到 AI-ready 的要求，因此，在数据通用规范中将明确规定，在具体标准中必须本着应收尽收原则，收集足够的元数据。在目前阶段，由于数据/元数据产生/收集过程使用的软、硬件没有考虑到应收尽收的需求，要完成这样的动作必然伴随着大量的手工记录与录入，致使数据管理占用大量时间与精力，实施者不胜其烦，不可避免地产生懈怠甚至抵触情绪。解决这一矛盾的关键在于尽快完成数据标准化，并将标准规则贯彻于软、硬件的配置中。随高通量实验与计算技术的发展，数据产生/收集过程必将实现全面自动化，这个问题也将必然逐步弱化直至消失。为此，一些基于实验设备或计算软件数据产生特点的工作流控制软件系统被开发出来，比如美国 NIST 开发了一套电子显微镜实验室信息管理系统—NexusLIMS[83]，可以将用户使用 Nexus 电子显微镜时间段内所产生的所有数据和元数据，都打包到一个表示实验快照的结构化文本文档中，实现所有原始研究数据自动备份和归档存储，同时构建了一个基于网络的门户网站，用户可按日期、用户、仪器、样本或任何其他元数据参数搜索访问之前的实验记录。计算材料领域由于其天然的规范化和数字化特征，也开发了多个围绕材料计算而开发的自动化数据工作流管理软件，包括 Fireworks[84]、AFLOW Π[85]、Atomate[86]，AiiDA[87]等，可实现计算数据的自动化采集和存储管理，在数据完整收集上具有相对优势。进一步地，对于通用的制备、表征、计算技术、方法、流程应建立统一的数据模板，即数据标准，使得这些数据可以方便地共享。

自《通则》发布以来，一系列围绕特定研究方法的示范性数据标准工作也正在积极开展。在“材料基因工程关键技术与支撑平台”国家重点专项的支持下，依据《材料基因工程数据通则》要求，构建了离子束溅射沉积样品元数据标准化模板，并建设在国家材料基因工程数据汇交与管理平台（http://nmdms.ustb.edu.cn），已被用于日常的科研数据管理中；在“云南省稀贵金属材料基因工程”重大科技专项支持下，围绕稀贵金属材料研究过程的数据标准化工作正在进行中，并在此基础上构建了一个大型的贵金属材料专业数据库（http://ipm-int.matclouds.com）。基于 DFT 方法的材料热性能计算数据的标准业已完成[88]。此外，结合高速列车车轮车轴产业化国家重点工程与综合领域共同制定了若干大尺寸构件全域高通量原位统计映射表征技术标准，以材料基因工程创新方法为评价相关材料构件的质量提供了科学支撑，现已申请立项 13 项，提出立项计划 30 多项。相关工作涵盖了数据的产生、采集、存储、共享和利用等环节，正在中国材料与试验团体标准委员会材料基因工程领域委员会成员单位中积极推进。CSTM 标准系统将确保材料基因工程研究活动及其成果的具有引领性、规范性、准确性、高效性和可复现性，而材料基因工程标准化的创新驱动，必将为材料产业高质量发展提供强有力的支撑。

2.1.3 材料数据标准体系

完整的 AI-ready 材料数据生态需要通过构建完整的数据标准体系来保证。《通则》为 AI-ready 材料数据的标准化建立了基点、指明了方向，也被用于更广泛意义上的材料数据标准设立所遵循的基本原则[80][81][82][88]。材料数据纷繁复杂，以《通则》为核心的数据标准化工作采取了一种自上而下与自下而上相结合的工作模式。首先，从顶层设计出发提出一套全面覆盖材料数据相关的方方面面问题的标准体系构架，对需要建立的标准进行了整体规划。在实操中依据标准体系框架，发动各方面专家，发挥各自专业特长，以《通则》为核心指导原则，从具体问题入手，逐步建立各类数据标准细则。材料基因工程数据标准体系框架如图 3 所示，从内容上可以划分为五个板块。

- 基础通用标准对材料数据的通用性要求进行明确。其中《通则》对材料数据的标准化工作目标、内容提供总体设计和规划。材料基因工程术语标准、数据标识标准、数据通用规范等标准，将《通则》的对数据的各项通用要求具体化，如前所述，分别为各类研究方法的数据标准制定提供权威术语、标识方法、标准化流程与方法参考，以更具体的服务、指导数据标准化工作的整体性建设，目前这三项通用标准正处于审核修订过程[80][81][82]。
- 实验数据和计算数据是有关材料数据产生的两个板块。相应的标准从材料数据生产者的角度出发，规定各种实验或计算方法产生的数据条目中应包含的内容。在具体执行上，需要重点关注三个方面：数据分类、标准建设粒度和标准化内容。首先，依据《通则》对材料数据的分类，按照实验制备/计算（虚

拟）制备、实验表征/计算表征、数据分析几种数据产生过程，将数据划分为样品信息、原始数据和衍生数据三类。其次，每件标准以可独立存在的数据产生动作（样品制备/表征/计算/数据处理）为条目主题，以该动作（样品制备/表征/计算/数据处理）所采用的具体方法为载体。例如针对“物理气相沉积方法（PVD）”制备薄膜样品过程，建立相应的“物理气相沉积（PVD）薄膜样品信息元数据标准”；针对“X 射线衍射分析（XRD）”表征，建立“XRD 表征元数据标准”；针对“XRD 数据物相分析”建立相应的“XRD 物相分析衍生元数据标准”。计算数据标准实例如 VASP 结构优化计算元数据标准（虚拟样品）、VASP 力常数计算元数据标准（虚拟表征）等。再者，标准的内容则是以数据产出动作过程为描述对象构建标准化的元数据模式。高通量实验与计算数据的标准除包含相应的样品制备/表征/计算/数据处理基本技术的规定外，还应反映高通量技术的特点。

- 数据应用标准板块包括一系列从材料数据在研究中应用角度出发，根据不同材料细分领域所关注的材料性质、参数来建立的标准化应用数据集元数据模式。比如针对低合金高强钢研究人们通常关注其关键成分、力学性能、组织结构、加工工艺等参数。领域专家根据多年经验，构建包括该材料常用特性的元数据模式，并形成领域共识，使其成为“低合金高强钢应用元数据标准”。数据应用标准依据材料类型划分粒度，为使用者提供一种专家经验的视角。
- 数据技术标准板块是从计算机科学出发，为材料数据标准在数据的存储、交互、挖掘、质量控制、数据安全等方面建立共识性协议、规范、标准，为数据在机器层面的一致性管理和互操作性提供信息技术保障，相关工作正在中国材料与试验团体标准委员会材料基因工程领域委员会成员单位中积极推进。

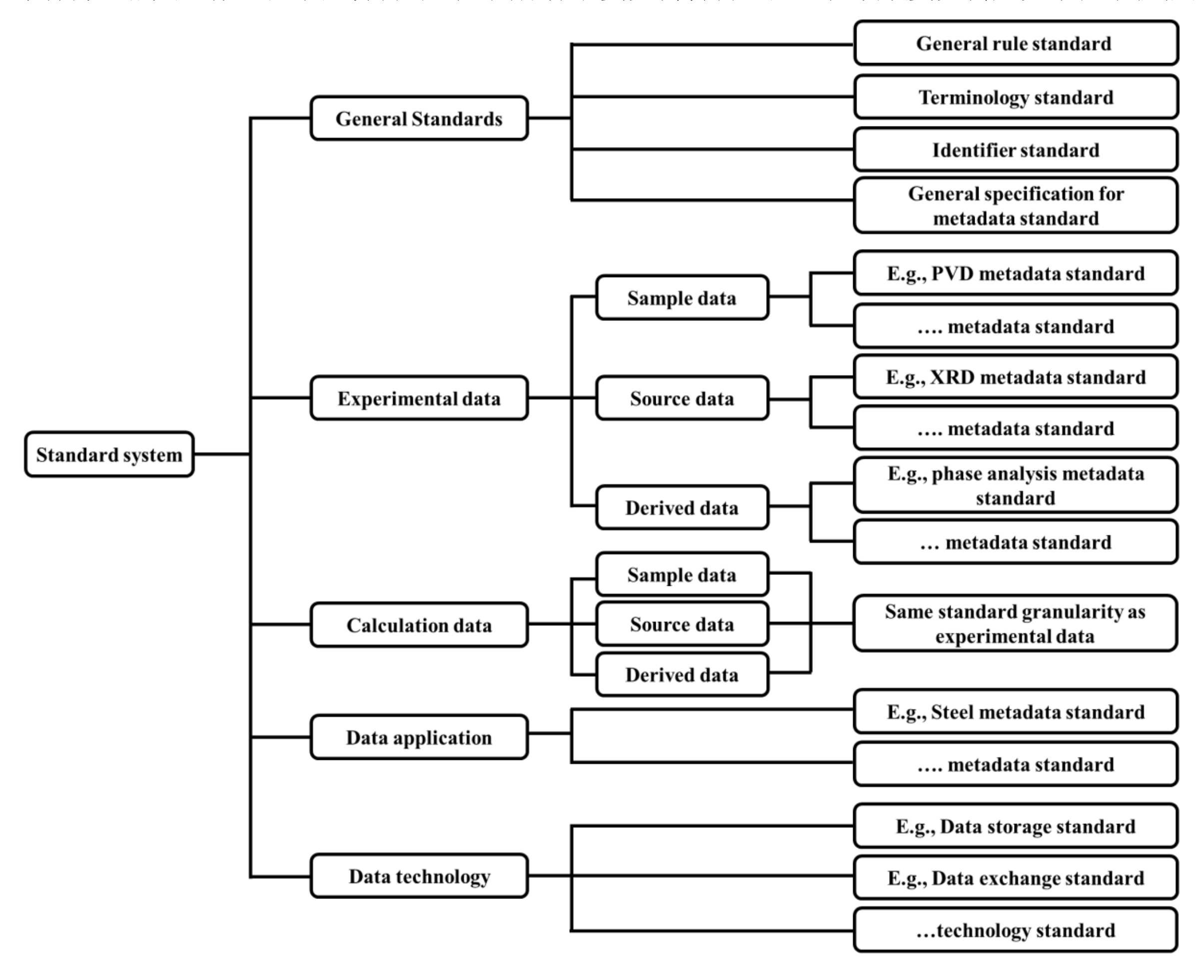


图 3 材料基因工程数据标准化工作框架示意图

Fig.3 Schematic diagram of the framework for standardization of materials genome engineering data

2.2 AI-Ready 数据基础设施

数据标准化的实施为构建完整可再用和可共享的规范化单条数据提供了治理方案，也为海量、特征全

面、均衡分布材料数据集的广泛构建奠定了基础。现有的材料研究基础设施是基于当前的需要而设计和开发的，产出的数据无论在量与质上，均与 AI 要求相差甚远。因此，AI-ready 数据的获得需要与之相符的新型材料数据基础设施予以支撑。新型材料创新基础设施将以数据为核心，AI 为关键词，由数据平台、高通量实验平台和高通量计算平台三部分组成。数据平台包括基于 AI 方法的软件工具库，与 AI-ready 的数据库；高通量实验与高通量计算平台作为数据生产来源，为快速获取大量数据提供了有效途径。这样，材料基因工程的 3 个技术要素实现了内在的协同，形成了缺一不可的深度融合关系。

构建 AI-ready 新型数据基础设施的相关技术包括了数据的高通量实验技术、数据的自动化采集存储技术、高通量计算技术、数据标准体系、数据语义和结构的标准化存储、数据的统一标识和网络访问获取等。数据标准化随每条数据渗透在其中的每一环节。通过综合运用这些技术，实现 AI-ready 数据产生、收集、存储、处理、交换、共享、使用、分析和网络协作的全链条综合基础能力[16]。

基于上述考虑，Wang 等[91]提出了“数据工厂”的概念模型，即在理想条件下，AI-ready 数据应产生于一个像工业生产线一样以标准化方式批量生产数据的专用设施平台。图 4 为数据工厂的概念图。概念图中央是数据工厂的数据设施。图 4 右翼为实验数据工厂，它可以是基于大型科学设施（如同步加速器光源、中子源等）的大规模、系统性的高通量综合制备与表征平台设施，集成一系列原位制备和多参数表征手段，能够产生包括力学、电气、光学、热学、磁学和声学特征及性能等多参量数据，理想情况下，所有性能测量都在同一样品上实时原位地进行。图 4 左翼展示了计算数据工厂的概念，它实质上是一个拥有各种高通量计算软硬件的计算中心，通过密度泛函理论、分子动力学、CALPHAD 方法、相场模拟、有限元分析等多种方法，配备有高通量计算工作流程，有能力生成从原子尺度到宏观尺度的大批量综合计算数据。数据工厂可以在同一地点集中建立，也可以由一组虚拟链接站点组成的分布式平台构成。

“数据工厂”将直接回应 AI-ready 对材料数据的各方面需求：自动化、不间断流水线式的数据采集存储方式为海量数据的产生提供了保证；公共数据生产设施弱化了研究者通常所带有的强烈目的性，使特征参数分布更为均衡；高通量的产生方式有利于获得具有更好的系统性、一致性的数据；综合的观测指标为人工智能对未知规律的探索提供了巨大的特征空间。数据标准可以方便地实施于数据工厂，使数据的采集、存储和管理数据都按照统一的方式进行，保证了 FAIR 原则在任何一条数据得到满足。同时，由于实现了自动化与标准化，以“应收尽收”原则收集大量参数不再是负担。

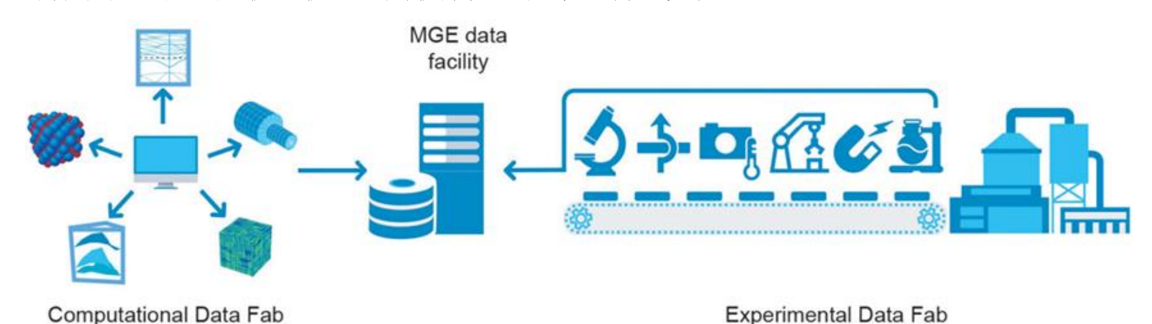


图 4 “数据工厂”概念图示——一个能够以标准化方式批量生产数据的专用设施，如同一条工业生产线[91]

Fig.4 Conceptualization of Data Fab—a dedicated facility capable of mass production of data in a standardized manner, just like an industrial production line[91]

“数据工厂”的出现将给数据生产带来一系列重大变革。首先，为了更广泛的长远的目标，综合、均衡的材料数据集将被大规模地有意识地产生，而不再局限于作为分散的具有特定目的的实验或计算的副产物；其二，数据标准的全面实施保证参数的完整性和数据的可共享性，使每条数据可用度和可用范围大幅提升；第三，“数据工厂”将数据产生由个体活动转变为有组织的社会活动。第四，这种有组织的努力将把数据的社会属性从私有财产转变为公共资源。其结果将带来材料数据数量和质量的全面提升，数据共享将变得更加简单，社会总成本也将降低。这种新型的数据产生方式是材料科学的革命性变化。

“数据工厂”概念模型反映了材料创新基础设施的最新发展趋势。在美国白宫国家科学技术委员会于 2021 年 11 月发布的最新“材料基因组计划战略规划”[17]中对材料创新基础设施作了重点部署，提出连接、创建和加强计算工具、实验工具及数据存储共享软件框架等要素，建设国家材料数据共享网络，将其整合

为一个统一化的材料研究连续体，从而扩大 MGI 覆盖范围，提高研究资源的易得性；在这个统一的数据网络架构基础上，以构建 AI-ready 数据为目标，利用和加强材料创新基础设施，使人工智能方法的应用大大加快材料的研发。

目前国际上已开发了一系列基于高通量计算平台或计算“数据工厂”的数据库平台。由麻省理工学院和劳伦斯·伯克利国家实验室联合开发的 Materials Project[92]，依托美国国家能源研究科学计算中心（National Energy Research Scientific Computing Center （NERSC））的超级计算集群，并借助其开发的 Fireworks 工作流软件和 Custodian 作业管理软件来自动管理计算及数据处理流程，建立了一个大型的材料第一性原理计算数据库，迄今已包括了超过 146000 种材料、24000 种分子、4000 多种电池材料等在内的系列计算性质数据，计算量达 1 亿 CPU 小时/年，并提供了多种检索、分析工具来帮助研究人员快速获取、分析数据（https://next-gen.materialsproject.org）。其它比较著名的高通量计算数据平台还有 Automatic Flow for Materials Discovery (AFLOW)[93]、Open Quantum Materials Database (OQMD)[94]、Novel Materials Discovery (NOMAD)[95]和 MatCloud[96]等。值得注意的是，这些基础设施在数据管理和存储时采用了各自独特的方式，相互之间并没有遵循同一标准，在多源数据整合为 AI-ready 数据时存在诸多不便[62]。OPTIMADE[79]联盟发布的通用 API 所支持的数据基础设施包括了 AFLOW、Materials Project、NOMAD、OQMD、Materials Cloud[97]等，通过 OPTIMADE API 可以在这些物理位置分布不同的材料数据基础设施实现跨库检索，体现出了“数据工厂”分布式建设、虚拟链接的特点。

与计算相比，具有“数据工厂”特点的实验数据大型数据库平台目前还较少， High Throughput Experimental Materials Database（HTEM DB）[98]是其中为数不多的典型代表，HTEM DB 由美国国家可再生能源实验室(National Renewable Energy Laboratory，NREL)基于其开展的物理气相沉积（PVD）组合薄膜样品的高通量制备和表征实验数据而建设，并开发了 LIMS 材料实验信息管理系统，负责自动收集、索引和归档实验数据，目前公共版本涵盖了 82000 余个采用物理气相沉积合成的各种薄膜材料样品（氧化物、氮化物、硫化物、磷化物、金属间化合物）的成分（55000+）、结构（65000+）、光学（46000+）和电学特性数据（19000+），同时提供了用户界面供研究者查询检索，并可通过提供的应用程序编程接口(Application programming interface，API)获取更多数据来进行数据挖掘和分析（https://htem.nrel.gov）。

3 结语

数据驱动模式为材料科学研究带来了颠覆性发展机会，数据的价值正在从辅助作用向核心作用转移。传统范式下形成的离散分布、多源异构、小规模、无规范的数据无法与 AI 实现有效对接，制约了数据驱动效力在材料领域的发挥，面向 AI 的数据治理和新型数据基础设施建设成为材料领域必须面对的问题。本文由 AI 分析原理出发，系统提出了构建 AI-ready 的材料数据所应满足的条件：海量、全面、完整、均匀和可共享，以期为数据驱动研究从更广领域构建更多、更可用的材料数据提供基本参考依据和方向。

标准化是实现 AI-ready 材料数据的重要基础，也是全球共同关注的问题。欧美国家注重与既有数据相匹配，着力通过建立整套材料科学本体，改善多源异构数据的可互操作性，但这种元数据协调方式仍需开发数据转换器和共享数据模式。我国通过建立《材料基因工程数据通则》重新定义了 AI-ready 材料数据的构建原则。基于《通则》核心理念提出的材料数据标准化框架体系，为 AI-ready 的材料数据生态的构建提供一套具体化的数据治理方案。不论采取何种方式，材料数据的标准化势在必行，但任重道远。

“数据工厂”新型数据基础设施是全面构建 AI-ready 数据库的理想场所，将为材料研究领域持续不断地提供海量、全面、完整、均匀、可共享的 AI-ready 标准化数据。当有一天“数据工厂”成为数据生产的主要形式时，数据驱动的潜力将有望真正得到释放。

作者贡献：汪洪，张澜庭:研究命题的提出；路勇超：研究方案设计，研究资料搜集，论文起草；余宁：文章修改建议；汪洪，张澜庭：文章质量控制，论证；路勇超，汪洪：负责最终修订版本。

参考文献

[1] White House Office of Science and Technology Policy. Materials genome initiative for global competitiveness [EB/OL]. (2011-06). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf
[2] Seventh framework programme of the European Community (EU FP7). ACCMET - Accelerated Metallurgy - the accelerated discovery of alloy formulations using combinatorial principles [EB/OL]. (2011-06). https://cordis.europa.eu/project/id/263206/reporting
[3] The Materials Science and Engineering Expert Committee (MatSEEC) of the European Science Foundation(EFS). Metallurgy Europe: a renaissance programme for 2012–2022 [R]. Strasbourg: EFS, 2012
[4] The Center for “Materials research by Information Integration” (CMI2) of MaDIS, NIMS “Materials research by Information Integration” Initiative (MI2I) [EB/OL]. (2015-07). https://www.nims.go.jp/MII-I/en/about/index_m.html
[5] Wang H, Xiang Y, Xiang X D, et al. Materials genome enables research and development revolution [J]. Sci. Technol. Rev., 2015, 33(10): 13
(汪洪，向勇，项晓东等. 材料基因组——材料研发新模式 [J]. 科技导报, 2015, 33(10):13)
[6] Su Y J, Fu H D, Bai Y, et al. Progress in Materials Genome Engineering in China [J]. Acta. Metall. Sin., 2020, 56(10): 1313
(宿彦京，付华栋，白洋等. 中国材料基因工程研究进展 [J]. 金属学报, 2020, 56(10): 1313)
[7] Wang H, Xiang X D, Zhang L T. Data + AI: The core of materials genomic engineering [J]. Sci. Technol. Rev., 2018, 36(14): 15
(汪洪，项晓东，张澜庭. 数据+人工智能是材料基因工程的核心 [J]. 科技导报, 2018, 36(14): 15)
[8] Lian S Y. Introduction to artificial intelligence [M]. Beijing: Tsinghua University Press, 2020: 3
(廉师友. 人工智能导论 [M]. 北京：清华大学出版社, 2020: 3)
[9] Liu F Q. Artificial intelligence [M]. Beijing: China machine Press, 2011: 1
(刘凤岐. 人工智能 [M]. 北京：机械工业出版社, 2011: 1)
[10] Kaplan A, Haenlein M. Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence [J]. Bus. Horiz., 2019, 62(1): 15
[11] Warren J A. The materials genome initiative and artificial intelligence [J]. MRS Bull., 2018, 43(6): 452
[12] Murdock R J, Kauwe S K, Wang A Y T, et al. Is domain knowledge necessary for machine learning materials properties? [J]. Integr. Mater. Manuf. Innov., 2020, 9(3): 221-227.
[13] Masood H, Toe C Y, Teoh W Y, et al. Machine learning for accelerated discovery of solar photocatalysts [J]. ACS Catal., 2019, 9(12): 11774-11787.
[14] Childs C M, Washburn N R. Embedding domain knowledge for machine learning of complex material systems [J]. MRS Commun., 2019, 9(3): 806-820.
[15] Liu Z H，Wang T Y. Data governance [M]. Beijing: Party School of the CPC Central Committee Press, 2021: 1
(李振华，王同益. 数据治理 [M].北京：中共中央党校出版社, 2021: 1)
[16] Fagnan K, Nashed Y, Perdue G, et al. Data and models: a framework for advancing AI in science [R]. United States: USDOE Office of Science (SC), 2019
[17] National Science and Technology Council, Committee on Technology and Subcommittee on the MGI Initiative. Materials genome initiative strategic plan [EB/OL]. (2021-11). https://www.mgi.gov/sites/default/files/documents/MGI-2021-Strategic-Plan.pdf
[18] Ghahramani Z. Probabilistic machine learning and artificial intelligence [J]. Nature, 2015, 521(7553): 452
[19] Schmidt J, Shi J, Borlido P, et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning [J]. Chem. Mat., 2017, 29(12): 5090
[20] Lee J, Seko A, Shitara K, et al. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques [J]. Phys. Rev. B, 2016, 93(11): 115104.
[21] Zhou J, Hong X, Jin P. Information fusion for multi-source material data: progress and challenges [J]. Appl. Sci, 2019, 9(17): 3473
[22] Kalidindi S R, De Graef M. Materials data science: current status and future outlook[J]. Ann. Rev. Mater. Res., 2015, 45: 171

[23] Schmidt J, Marques M R G, Botti S, et al. Recent advances and applications of machine learning in solid-state materials science [J]. npj. Comput. Mater., 2019, 5(1): 1

[24] Schmidt J, Shi J, Borlido P, et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning [J]. Chem. Mat., 2017, 29(12): 5090

[25] De Jong M, Chen W, Notestine R, et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds [J]. Sci. Rep., 2016, 6(1): 1

[26] Kim J, Kang D, Kim S, et al. Catalyze materials science with machine learning [J]. ACS Mater. Lett., 2021, 3(8): 1151-1171.

[27] Liu Y H, Hu Z H, Suo Z G, et al. High-throughput experiments facilitate materials innovation: a review [J]. Sci. China-Technol. Sci., 2019, 62(4): 521

[28] Curtarolo S, Hart G L W, Nardelli M B, et al. The high-throughput highway to computational materials design [J]. Nat. Mater., 2013, 12(3): 191

[29] Hattrick-Simpers J R, Gregoire J M, Kusne A G. Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge [J]. APL. Mater., 2016, 4(5): 053211

[30] Xiang X, Sun X, Briceño G, Lou Y, WangK, Chang H, et al. A combinatorial approach to materials discovery [J]. Science 1995,268(5218):1738

[31] Ceder G, Persson K. The stuff of dreams [J]. Sci. Am., 2013, 309(6): 36

[32] Zhang X, Chen A, Zhou Z. High-throughput computational screening of layered and two-dimensional materials [J]. Wires Comput. Mol. Sci., 2019, 9(1): e1385.

[33] Brunin G, Ricci F, Ha V A, et al. Transparent conducting materials discovery using high-throughput computing [J]. npj. Comput. Mater., 2019, 5(1): 1

[34] Kononova O, He T, Huo H, et al. Opportunities and challenges of text mining in materials research [J]. Iscience., 2021, 24(3): 102155

[35] Blokhin E, Villars P. The PAULING FILE project and materials platform for data science: From big data toward materials genome [A]. Handbook of Materials Modeling-Methods: Theory and Modeling [C], Cham: Springer, 2020: 1837

[36] Swain M C, Cole J M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature [J]. J. Chem. Inf. Model., 2016, 56(10): 1894

[37] Court C J, Cole J M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction [J]. Sci. Data., 2018, 5(1): 1

[38] Huang S, Cole J M. A database of battery materials auto-generated using ChemDataExtractor [J]. Sci. Data., 2020, 7(1): 1

[39] Moosavi S M, Jablonka K M, Smit B. The role of machine learning in the understanding and design of materials [J]. J. Am. Chem. Soc., 2020, 142(48): 20273

[40] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials [J]. npj Comput. Mater., 2016, 2(1): 1

[41] Ghiringhelli L M, Baldauf C, Bereau T, et al. Shared metadata for data-centric materials science [J]. arXiv, 2022: 2205.14774

[42] Baker M. 1,500 scientists lift the lid on reproducibility [J]. Nature, 2016, 533(7604)

[43] Editorial. The importance and challenges of data sharing [J]. Nat. Nanotechnol., 2020(15):83

[44] Thelwall M, Kousha K. Figshare: a universal repository for academic resource sharing? [J]. Online Inf. Rev., 2016

[45] Dillen M, Groom Q, Agosti D, et al. Zenodo, An archive and publishing repository: a tale of two herbarium specimen pilot projects[J]. Biodivers. Inf. Sci. Stand, 2019 (2)

[46] White H, Carrier S, Thompson A, et al. The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment[A]. Proceedings of the International Conference on Dublin Core and Metadata Applications [C]. Göttingen: Universitätsverlag Göttingen, 2008:157

[47] Krawczyk B. Learning from imbalanced data: open challenges and future directions [J]. Prog. Artif. Intell., 2016, 5(4): 221

[48] Weissman J. Amazon created a hiring tool using aI it immediately started discriminating against women [J]. Slate, 2018

[49] Davenport T, Guha A, Grewal D, et al. How artificial intelligence will change the future of marketing [J]. J. Acad. Mark. Sci.,

2020, 48(1): 24

[50] Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments [J]. Nature, 2016, 533(7601): 73

[51] White House Office of Management and Budget. Federal data strategy 2020 action plan [EB/OL]. (2019-12). https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf

[52] European Commission. A European strategy for data [EB/OL]. (2020-2). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN

[53] General Office of the State Council of the People's Republic of China. Scientific Data Management Measures [EB/OL]. (2018-3). https://www.cgs.gov.cn/xwl/zfw/201804/W020180403526880358641.pdf
(中华人民共和国国务院办公厅.科学数据管理办法[EB/OL]. (2018-3). https://www.cgs.gov.cn/xwl/zfw/201804/W020180403526880358641.pdf)

[54] Kozlov M. NIH issues a seismic mandate: share data publicly [J]. Nature, 2022,602(7898):558

[55] Mauthner N S, Parry O. Open access digital data sharing: Principles, policies and practices [J]. Soc. Epistemol., 2013, 27(1): 47

[56] Tenopir C, Dalton E D, Allard S, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide [J]. PloS one, 2015, 10(8): e0134826

[57] Wilkinson M D, Dumontier M, Aalbersberg I J J, et al. The FAIR Guiding Principles for scientific data management and stewardship [J]. Sci. Data., 2016, 3(1): 1

[58] Berman F. The research data alliance--the first five years [EB/OL]. (2019). https://www.rd-alliance.org/sites/default/files/attachment/RDA%20RETROSPECTIVE%20FINAL%20-%20HDSR.pdf

[59] Mons B, Neylon C, Velterop J, et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud [J]. Inf. Serv. Use, 2017, 37(1): 49

[60] Liu J. Digital Object Identifier (DOI) and DOI Services: An Overview[J]. Libri, 2021, 71(4): 349

[61] State General Administration of the People's Republic of China for Quality Supervision and Inspection and Quarantine, Standardization Administration of China. GB/T 20000.1-2014 Guidelines for standardization—Part 1: Standardization and related actives—General vocabulary[S]. Beijing: Standards Press of China, 2014
(中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会. GB/T20000.1-2014，标准化工作指南第 1 部分：标准化和相关活动的通用术语 [S]. 北京：中国标准出版社, 2014)

[62] Himanen L, Geurts A, Foster A S, et al. Data - driven materials science: status, challenges, and perspectives [J]. Adv. Sci., 2019, 6(21): 1900808

[63] Scheffler M, Aeschlimann M, Albrecht M, et al. FAIR data enabling new horizons for materials research [J]. Nature, 2022, 604(7907): 635

[64] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods [J]. Data Knowl. Eng., 1998, 25(1-2): 161

[65] Yang T. Research on some key technologies of agricultural knowledge service based on ontology [D]. Shanghai: Fudan University, 2011
（杨涛. 基于本体的农业领域知识服务若干关键技术研究 [D]. 上海：复旦大学,2011.）

[66] Zhang X, Zhao C, Wang X. A survey on knowledge representation in materials science and engineering: An ontological perspective [J]. Comput. Ind., 2015, 73: 8

[67] Ghiringhelli L M, Carbogno C, Levchenko S, et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats [J]. npj Comput. Mater., 2017, 3(1): 1

[68] Tadmor E B, Elliott R S, Sethna J P, et al. The potential of atomistic simulations and the knowledgebase of interatomic models [J]. Jom, 2011, 63(7): 17.

[69] van der Vet P E, Speel P H, Mars N J I. The Plinius ontology of ceramic materials [A]. Eleventh European Conference on Artificial Intelligence (ECAI'94) Workshop on Comparison of Implemented Ontologies [C]. New York: John Wiley & Sons, 1994: 187

[70] Sainte Marie C, Iglesias Escudero M, Rosina P. The ONTORULE project: where ontology meets business rules [A].

International Conference on Web Reasoning and Rule Systems [C]. Berlin: Springer, 2011: 24
[71] Premkumar V, Krishnamurty S, Wileden J C, et al. A semantic knowledge management system for laminated composites [J]. Adv. Eng. Inform., 2014, 28(1): 91
[72] Michel K, Meredig B. Beyond bulk single crystals: a data format for all materials structure–property–processing relationships [J]. Mrs Bull., 2016, 41(8): 617
[73] Ashino T. Materials ontology: An infrastructure for exchanging materials information and knowledge [J]. Data Sci. J., 2010, 9: 54
[74] European Materials Modelling Council. EMMO: an Ontology for Applied Sciences [EB/OL]. (2017). https://emmc.info /emmo-info/
[75] Cheung K, Drennan J, Hunter J. Towards an Ontology for Data-driven Discovery of New Materials [A]. AAAI Spring Symposium: Semantic Scientific Knowledge Integration [C]. Menlo Park: The AAAI Press, 2008: 9
[76] Bhat M, Shah S, Das P, et al. Premλp: knowledge driven design of materials and engineering process [A]. ICoRD'13 international conference on research into design [C]. India: Springer, 2013: 1315
[77] Zhang X, Hu C, Li H. Semantic query on materials data based on mapping MATML to an OWL ontology [J]. Data Sci. J., 2009, 8: 1
[78] Ramakrishna S, Zhang T Y, Lu W C, et al. Materials informatics [J]. J. Intell. Manuf., 2019, 30(6): 2307
[79] Andersen C W, Armiento R, Blokhin E, et al. OPTIMADE, an API for exchanging materials data [J]. Sci. Data, 202 1, 8(1): 1
[80] China Standards of Testing and Materials (CSTM). Announcement on the establishment of the CSTM standard "materials genome terminology". [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/8a276faa-f242-4c3b-8bf1-4fb238af8ef3
(中国材料与试验团体标准委员会. 关于 CSTM 标准《材料基因工程术语》的立项公告. [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/8a276faa-f242-4c3b-8bf1-4fb238af8ef3
[81] China Standards of Testing and Materials (CSTM). Announcement on the establishment of the CSTM standard "Data Identifier Naming Method for Materials Genome Engineering". [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/b356a5f0-a75e-4671-8b7a-f95f95351ade
(中国材料与试验团体标准委员会. 关于 CSTM 标准《材料基因工程数据标识符命名方法》的立项公告. [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/b356a5f0-a75e-4671-8b7a-f95f95351ade
[82] China Standards of Testing and Materials (CSTM). Announcement on the establishment of the CSTM standard "General Metadata Specification for Materials Genome Engineering Data". [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/69bfbb17-e88e-481a-bd16-756cdbc969cd
(中国材料与试验团体标准委员会. 关于 CSTM 标准《材料基因工程数据通用元数据规范》的立项公告. [EB/OL]. (2021-11), http://www.cstm.com.cn/article/details/69bfbb17-e88e-481a-bd16-756cdbc969cd
[83] Taillon J A, Bina T F, Plante R L, et al. NexusLIMS: A laboratory information management system for shared-use electron microscopy facilities [J]. Microsc. microanal., 2021, 27(3): 511
[84] Jain A, Ong S P, Chen W, et al. FireWorks: A dynamic workflow system designed for high‐throughput applications [J]. Concurr. Comput., 2015, 27(17): 5037
[85] Supka A R, Lyons T E, Liyanage L, et al. AFLOWπ: A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians [J]. Computational Materials Science, 2017, 136: 76
[86] Mathew K, Montoya J H, Faghaninia A, et al. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows [J]. Comput. Mater. Sci., 2017, 139: 140
[87] Pizzi G, Cepellotti A, Sabatini R, et al. AiiDA: automated interactive infrastructure and database for computational science [J]. Comput. Mater. Sci., 2016, 111: 218
[88] Rao Y C, Lu Y C, Ju S H, et al. Metadata standard for phonon thermal conductivity by first-principles calculation [J]. J. Mater. Inf., 2022, (In-revision)
[89] Chinese Society fot Testing & Materials, Zhongguancun. T/CSTM00120-2019 General rule for materials genome engineering

data [S]. Beijing: Metallurgical Industry Press, 2019
(中关村材料试验技术联盟. T/CSTM00120-2019,材料基因工程数据通则 [S]. 北京: 冶金工业出版社, 2019)
[90] State General Administration of the People's Republic of China for Quality Supervision and Inspection and Quarantine, Standardization Administration of China. GB/T232843-2016 Science and technology resource identification [S]. Beijing: Standards Press of China, 2016
(中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会. GB/T232843-2016, 科技资源标识 [S]. 北京: 中国标准出版社, 2016)
[91] Wang H, Xiang X D, Zhang L T. On the data-driven materials innovation infrastructure [J]. Engineering., 2020, 6(6): 609
[92] Jain A, Ong S P, Hautier G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation [J]. APL. Mater., 2013, 1(1): 011002
[93] Curtarolo S, Setyawan W, Hart G L W, et al. AFLOW: An automatic framework for high-throughput materials discovery [J]. Comput. Mater. Sci., 2012, 58: 218
[94] Kirklin S, Saal J E, Meredig B, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies [J]. npj Comput. Mater., 2015, 1(1): 1
[95] Draxl C, Scheffler M. NOMAD: The FAIR concept for big data-driven materials science [J]. MRS. Bull., 2018, 43(9): 676
[96] Yang X, Wang Z, Zhao X, et al. MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources [J]. Comput. Mater. Sci., 2018, 146: 319
[97] Talirz L, Kumbhar S, Passaro E, et al. Materials Cloud, a platform for open computational science [J]. Sci. Data, 2020, 7(1): 1
[98] Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials [J]. Sci. Data, 2018, 5(1): 1